Power Usage of Production Supercomputers and Production Workloads

Scott Pakin, Curtis Storlie, Michael Lang, Robert E. Fields III, Eloy E. Romero, Jr., Craig Idler,
Sarah Michalak, Hugh Greenberg, Josip Loncaric, Randal Rheinheimer, Gary Grider, Joanne Wendelberger
Los Alamos National Laboratory
Los Alamos, New Mexico 87545

Abstract—**Power is becoming an increasingly important concern for large supercomputer centers. However, to date, there have been a dearth of studies of power usage "in the wild"—on production supercomputers running production workloads. In this paper, we present the initial results of an effort to characterize the power usage of three Top500 supercomputers at Los Alamos National Laboratory: Cielo, Roadrunner, and Luna (#6, #10, and #45, respectively, on the November 2011 Top500 list). Power measurements taken both at the switchboard level and within the compute racks are presented and discussed.**

I. Introduction

A major challenge of exascale computing is to deliver a thousandfold increase in performance while only slightly increasing power consumption over current petascale systems [1], [2]. While there has been much research on increasing power efficiency within various components of a supercomputing system—processors, interconnection networks, system software, programming models, algorithms, and applications—and numerous controlled studies, there is comparatively little data available describing how much power a supercomputer draws while running a production scientific workload. The goal of this paper is to fill that gap by presenting power data measured at the main supercomputer data center at Los Alamos National Laboratory (LANL) and on three of the world's fastest supercomputers, based on the annual Top500 list of supercomputer performance [3].

In particular, we present full-system power data measured since inception of two production supercomputers, Cielo and Roadrunner, and one preproduction supercomputer, Luna. Combined, these systems represent a total of just under 13,500 nodes, making this, to our knowledge, the largest study of power drawn during a real workload. For Luna, we further include some controlled studies to analyze the extremes of that system's power usage and examine the discrepancies between measuring power at the switchboard level and at the sub-rack level.

Our findings are that power variability differs substantially across architectures; from a power perspective, real scientific workloads bear little in common with the LINPACK benchmark (not too surprisingly); the difference between worst-case and average-case power draws indicates that supercomputing data centers may contain a fair amount of "trapped capacity" in their power systems, more if power capping can be implemented on a full-system basis; job schedulers theoretically have the potential to increase trapped capacity even further by pairing jobs of different power envelopes; and energy savings are unlikely to be achieved merely by frequency and voltage scaling, given how supercomputers are currently run.

We anticipate that this paper will assist future power studies that require knowledge of real-world supercomputer power data to drive their approach and solutions.

The remainder of the paper is organized as follows. We discuss the most relevant pieces of related work in Section II. Section III describes LANL's data center in terms of its power characteristics and the main supercomputers it hosts. The main section of the paper is Section IV, where we present our power measurements and associated analyses. Section V briefly describes some prospects for follow-on research. Finally, we draw some conclusions from our findings in Section VI.

II. Related Work

Fan, Weber, and Barroso quantify the power usage of three workloads that run at one of Google's large data centers: Google Web search, GMail, and various offline MapReduce jobs [4]. As in our work, they examine power characteristics at the rack, sub-cluster, and full-cluster levels. The key characteristic that distinguishes our work from theirs is that we focus on a production scientific workload in which applications tend to be more tightly coupled than search and email services and MapReduce jobs. Scientific applications generally include substantial communication within sets of processes/nodes, and the workload as a whole tends to allocate and free large numbers of nodes at once. This can incur sudden power changes and therefore requires great care in implementing power capping. Also, unlike Google's large data centers, LANL's supercomputers run at fairly constant job load over time, limiting the usefulness of the Google paper's studies of reducing power during off-peak times. LANL's data center has no off-peak times. Although incidental, our paper presents data that covers twice as many nodes and for twice the duration as Fan, Weber, and Barroso's work.

In the context of scientific computing, Laros et al. have performed a number of large-scale power studies on two Cray XT supercomputers [5]. They quantified the power usage of a number of representative scientific applications with different CPU frequencies and by introducing power savings in the network by scaling back communication bandwidth. The main difference between our work and Laros et al.'s is that they performed controlled studies of individual applications while our study represents a production workload with many applications of different sizes running concurrently in a space-shared manner across entire systems.

III. Facility

Figure 1 illustrates our power-monitoring infrastructure. 13,200V enter the data center and are converted to 3-phase, 480V power to feed the substations. The substations transmit power through a rotary uninterruptible power supply (RUPS) to a number of switchboards, each of which feeds multiple power distribution units (PDUs) on the machine-room floor. These convert the power into 3-phase, 208V or 480V power for distribution to the compute racks. An assumption the facilities engineers make is that no PDU will ever exceed 80% load (e.g., 200 kVA for a 240 kVA PDU). Otherwise, the unit runs the risk of observing a voltage sag, tripping the circuit breaker, and cutting power to the associated racks (and thereby making many users quite unhappy).

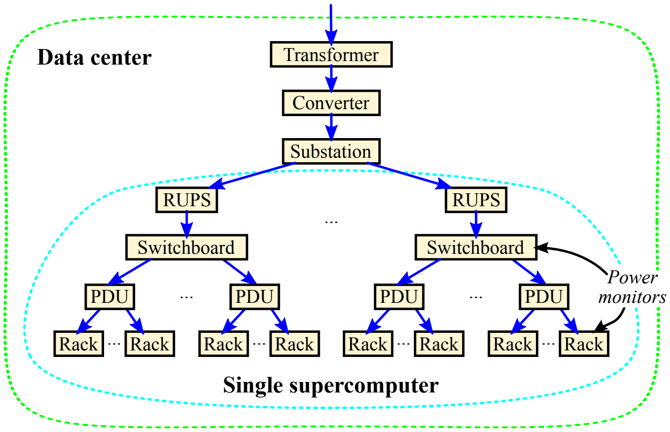


Fig. 1: Power-monitoring infrastructure

Our data center can deliver a aggregate of 19.2 MW of power. From January through April 2012, the facility has generally been running at about half of peak capacity and has averaged a power usage effectiveness (PUE)—the ratio of total power to IT equipment power [6]—of 1.41, which, while far from the state of the art in efficient data-center design, is nevertheless substantially better than the 1.86 averaged by the data centers in Greenberg et al.'s study [7].

Power can be monitored at each switchboard and at each rack. In fact, two of our supercomputers each provide finer than rack-resolution monitoring: Roadrunner supports intra-node monitoring, and Luna supports monitoring at the "shelf" (10-node) level. Switchboard monitoring is automatically logged and stored. Originally, logging was performed at 15-minute intervals. In late February/early March 2012 we increased the logging rate to 1-minute intervals to help correlate the switchboard readings with other power monitors. Currently, rack and sub-rack monitoring is done only on demand, by explicitly polling the monitoring devices. We therefore have comparatively little data at this level.

An important feature of the way our data center is configured is that each switchboard is associated with a single supercomputer. Consequently, we can measure power independently for each supercomputer, a capability that would not otherwise be possible. Note, however, that cooling and storage (parallel filesystems) are not associated with any particular supercomputer and reside on separate switchboards.

Table I lists some key characteristics of the systems we used in our study. **Roadrunner** was the world's first petascale system [8]. It uses a hybrid architecture with AMD Opteron CPUs and IBM Cell processors as computational accelerators. Each node comprises a total of 40 cores: two sockets of dual-core Opteron and four sockets of Cell, with each Cell socket containing one general-purpose control processor and eight vector processors [9]. Nodes are connected by a dual-data-rate (DDR) InfiniBand [10] fat tree from Mellanox. Due to the Cell's high peak performance per watt, Roadrunner ranked sixth on the second Green500 list [11] even though its aggregate performance was substantially higher than virtually every other system on the list.

Cielo is a large Cray XE6 system [12]. Each node contains two sockets of 8-core AMD Magny-Cours CPUs, and nodes are connected with a Cray Gemini network [13], organized as a 3-D torus. What makes Cielo interesting from a power perspective is that the communication fabric is integrated into the nodes, not separated as it is in the other two systems in Table I. Consequently, racks are more homogeneous in Cielo, while Roadrunner and Luna incorporate network switches at various points in the system. For example, Roadrunner's first-level InfiniBand switches are housed in 17 out of its 272 compute racks, and there are an additional 4 second-level InfiniBand switches separate from the compute racks but on a shared switchboard.

Luna is a significantly different system from both Roadrunner and Cielo. First, it is commodity cluster with no custom hardware [14]. Each node contains two sockets of 8-core Intel Sandy Bridge CPUs, and nodes are connected with a quad-data-rate (QDR) InfiniBand [10] fat tree from QLogic. Second, Luna is intended to run a large number of small jobs (tens to hundreds of nodes) rather than a small number of large jobs (thousands of nodes), as is the case for Roadrunner and Cielo. Finally, while Roadrunner and Cielo are running a production workload, at the time we gathered our data, Luna had not yet stabilized to the point where general users were allowed onto the system. (It has since joined Roadrunner and Cielo as a production resource.) Because of Luna's inchoate status, it was comparatively easy to reserve the entire machine for our work; our controlled studies are therefore all performed on Luna. Note that the Top500 data (rank and power consumption) for Luna shown in Table I is a bit misleading as it represents only about half the system. This was all that was installed at the time of the November 2011 Top500 submission due date.

Roadrunner, Cielo, and (soon) Luna are used almost exclusively for large-scale scientific simulations of national importance. These simulations help ensure the safety, security, reliability, and performance of the U.S. nuclear-weapons

TABLE I: Supercomputers used in this study

Machine	Top500 rank	Switch-boards	Racks	Nodes	Cores	Max. power/ rack (kW)	Max. power/ system (MW)	LINPACK power (MW)
Roadrunner	10	4	272	3,060	122,400	15	4.08	2.35
Cielo	6	5	96	8,892	142,272	54	5.18	3.98
Luna	45[a]	1	35	1,540	24,640	24	0.84	0.28[a]
Total	—	10	403	13,492	289,312	—	10.1	—

[a] The November 2011 LINPACK data represents an incomplete version of Luna containing only 14,080 cores.

stockpile without nuclear explosive testing [15]; they provide a comprehensive understanding of the threat posed by weapons of mass destruction; and they inform critical decisions related to the entire nuclear-weapons life cycle, from design to safe processes for dismantlement [16]. However, simulating macroscopic phenomena at the subatomic level leads to an unsatiable demand for computing resources. Typical applications run for long periods of time (often months of wall-clock time) on large numbers of processors and perform only occasional I/O (a few minutes every few hours) for checkpointing or writing out results [17]. However, in addition to large, scientific simulations, LANL supercomputers host a number of single-node jobs such as compilations, debugging sessions, interactive usage, and other tasks. Furthermore, it is common practice to run a large number of modest-sized parameter sweeps (tens of nodes) to identify likely parameters of interest then feed those parameters into a few long-running simulations that occupy many hundreds or thousands of nodes. In short, the workload running on LANL supercomputers is qualitatively different from that running on, say, Google's [4].

IV. Measurements and Analysis

In this section we consider two sets of power measurements. Section IV-A analyzes the power drawn by three supercomputers over a 16-month period running their normal workloads. Section IV-B reports the results of some controlled power experiments performed on one of these supercomputers in a single 10-hour block of time.

A. Production workloads

We begin by presenting the power drawn over time for each of Roadrunner, Cielo, and Luna. Figure 2 shows measured power in kilowatts over a date range from the start of 2011 to the end of April, 2012 for the three systems, and Table II summarizes the data. The following are some points of clarification:

- The y axis varies across the subfigures to more clearly show the measured power relative to each machine's theoretical peak power and the power it consumed while running the LINPACK benchmark, which is the metric that orders the Top500 list of supercomputer performance [18],
- Luna's LINPACK power is omitted because LINPACK data has not yet been reported for the complete system.
- Luna came online more recently than the other two systems so its power data extends over a lesser range.
- The small gap in data from the end of June to the beginning of July 2011 represents a complete machine-room shutdown as a precautious measure due to a major wildfire coming dangerously close to the Laboratory [19].
- The large gap in data from October to November is due to an upgrade of the switchboard-monitoring software that resulted in some data loss.

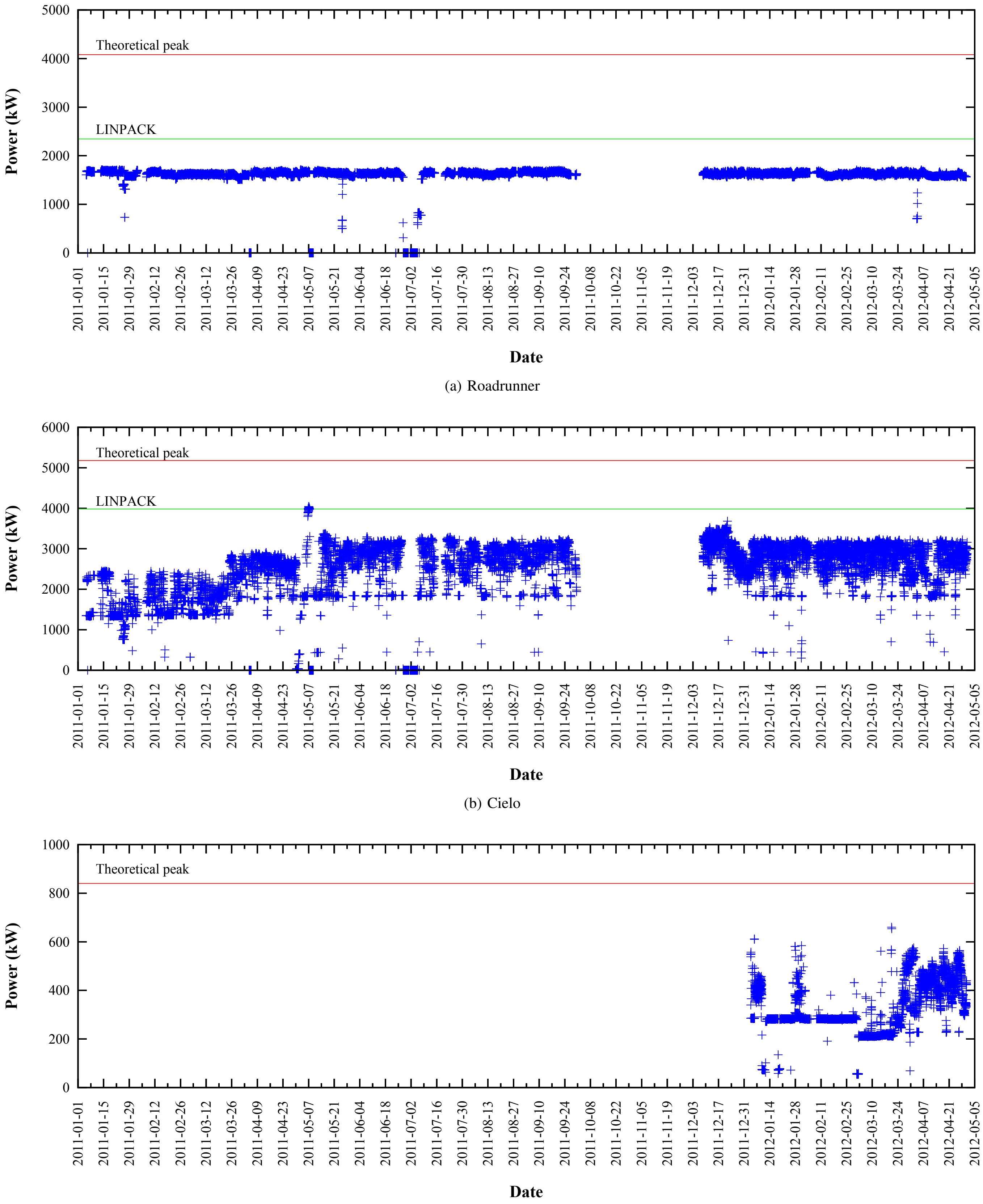
TABLE II: Descriptive statistics of supercomputer power usage

Statistic	Roadrunner (kW)	Cielo (kW)	Luna (kW)
Maximum	1,733	4,043	682
Median	1,632	2,988	352
Mean (μ)	1,623	2,839	327
Std. dev. (σ)	110	555	126

1) Initial observations: The following are some observations one can make from the data shown in Figure 2 and Table II. First, Roadrunner has the most consistent power draw of the three supercomputers, with a relative standard deviation of only 6.8% versus almost 20% for Cielo and almost 40% for Luna. To determine if Cielo's power variability is localized or spread across the entire system, we plotted the individual power contribution of each of Cielo's five switchboards. As Figure 3 indicates, all five switchboards observe similar fluctuations in power draw. Not all switchboards service the same number of nodes, which is why the bottom two curves (magenta and red) exhibit different average power draws from the top three curves (green, blue, and cyan).

The second obseration one can make from Figure 2 and Table II is that the scientific workload running on Roadrunner draws only 69.4% of the power that LINPACK draws, and the scientific workload running on Cielo draws only 75.1% of LINPACK power. These numbers indicate that from a power perspective, LINPACK is not particularly representative of the scientific workload normally executed at LANL.

As a third observation, one can quantify what is sometimes referred to as a supercomputer's "trapped capacity"—the difference between the infrastructure capacity allocated to a given system and the actual peak demand of that system. Going by the maximum power draw ever observed in the timeframe represented by the data, can consider Roadrunner to have 2,347 kW (58%), Cielo to have 1,137 kW (22%), and Luna to have 158 kW (19%) of trapped capacity. In practice, however, facilities engineers typically include a safety margin between the maximum power a system is expected to draw and the



(a) Roadrunner

(b) Cielo

(c) Luna

Fig. 2: Power draw over time of three LANL supercomputers

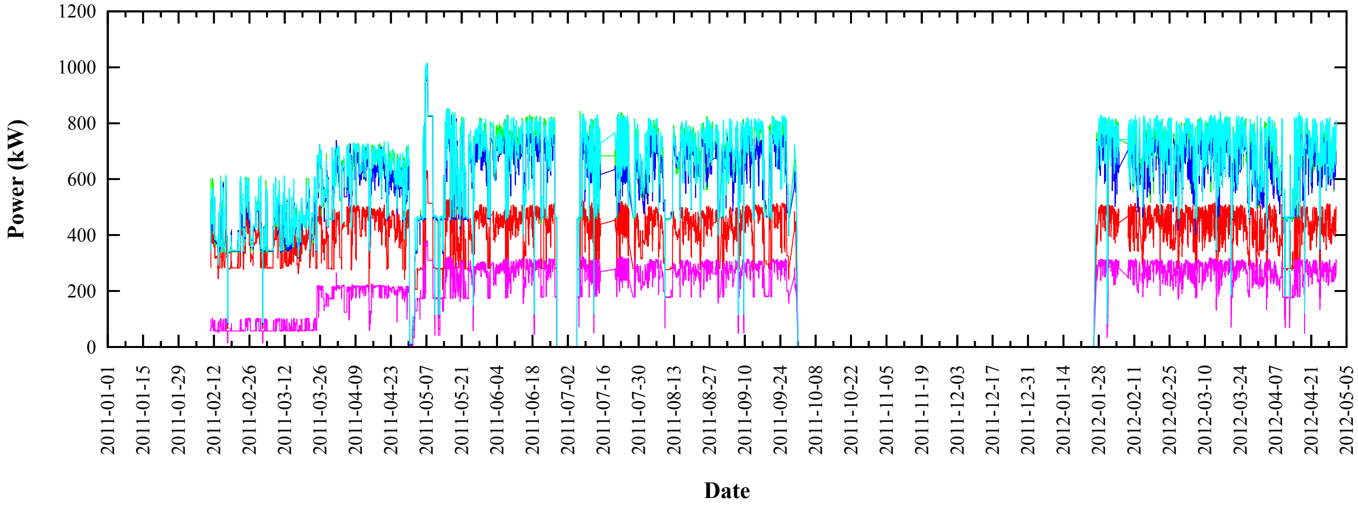


Fig. 3: Per-switchboard power data for each of Cielo's five switchboards

power available to that system to reduce the risk of tripping a circuit breaker, as mentioned in Section III.

We now examine our data center's trapped capacity in greater detail.

2) Trapped capacity: Table III quantifies the amount of trapped capacity available in LANL's data center. It assumes a safety margin of 20% below the maximum power listed in Table I and considers different definitions of peak system demand including the measured LINPACK value and the maximum power draw observed. Note, however, that these two values may not be distinct; either some applications do in fact have similar power characteristics to LINPACK or, more likely, our measurements include a few LINPACK runs.

Trapped capacity can be increased even further if one assumes the ability to explicitly limit the power a system draws by throttling frequency and voltage parameters, for example by using Intel's Node Manager [20]. Table III therefore also lists the available trapped capacity if power is capped to the mean power (μ) plus 1, 2, or 3 standard deviations (σ), enabling, respectively, an expected 68.3%, 95.5%, or 99.7% of the workload to run at full speed.

For example, Table I lists Luna's maximum power as 0.84 MW. Subtracting a 20% safety margin reduces this to a usable maximum of 672 kW. Table II lists the mean and standard deviation power draw of the programs running on Luna as 327 and 126 kW, respectively. If we require 95.5% of the workload to run at full speed, Luna will need $\mu + 2\sigma = 327 + 2 \times 126 = 579$ kW of power delivered to it. This leaves Luna with $672\,\text{kW} - 579\,\text{kW} = 93\,\text{kW}$ ($\sim$14% of 672 kW) of trapped capacity, which is what is listed in Table III.

The data in Table III indicate that there is a large potential for power improvements for Roadrunner but less for the other systems unless power capping is utilized (which, according to Laros et al.'s study, will degrade the performance of some applications more than others [5]). By "power improvements" we mean that more racks can be added without having to upgrade the facility's power infrastructure or that power costs can be reduced by delivering less power to the system.

TABLE III: Trapped capacity relative to various power maxima, assuming a safety margin of 20%

Assumed max. power		Roadrunner (kW)	Cielo (kW)	Luna (kW)
LINPACK		919 (28%)	164 (4%)	—
$\mu+\sigma$	(68.3%)	1,532 (47%)	750 (18%)	219 (33%)
$\mu+2\sigma$	(95.5%)	1,422 (44%)	195 (5%)	93 (14%)
$\mu+3\sigma$	(99.7%)	1,312 (40%)	0 (0%)	0 (0%)
Max. obs.	(100.0%)	1,531 (47%)	101 (2%)	0 (0%)

3) Job scheduling: It is possible to establish an upper bound on the potential gain in power effficiency that can be achieved by power-aware job scheduling. The maximum potential savings in peak power for the workload represented by Figure 2 would be hypothetically achieved by somehow arranging all processes so that each supercomputer was running constantly at the mean power draw for the period (i.e., with no variability in power draw). This is because all of the processes still would have to run, so the overall power used for the period is fixed. The way to minimize the maximum power draw would therefore be to have it remain constant at the mean level. Thus, the maximum potential reduction in peak power that can be attained by a perfect power-aware job scheduler (and assuming completely malleable applications) is $1 - \text{mean power}/\text{max. power}$. For the production supercomputers and production workloads used in our study, this implies a maximum power improvement of 6.3% on Roadrunner, 29.8% on Cielo, and 52.1% on Luna, according to Table II. Again, this is a crude upper bound that makes some unrealistic assumptions about application characteristics, but it does indicate that there may exist the potential for a job scheduler to improve the power usage of a large-scale scientific workload.

4) Energy usage: While most of our study focuses on power, energy—the integral of power over time—is also an important concern. We want to answer the following question:

> *For LANL's workload, if power is reduced to the bare minimum (i.e., idle power), how much slowdown in execution speed can the workload tolerate without increasing total energy?*

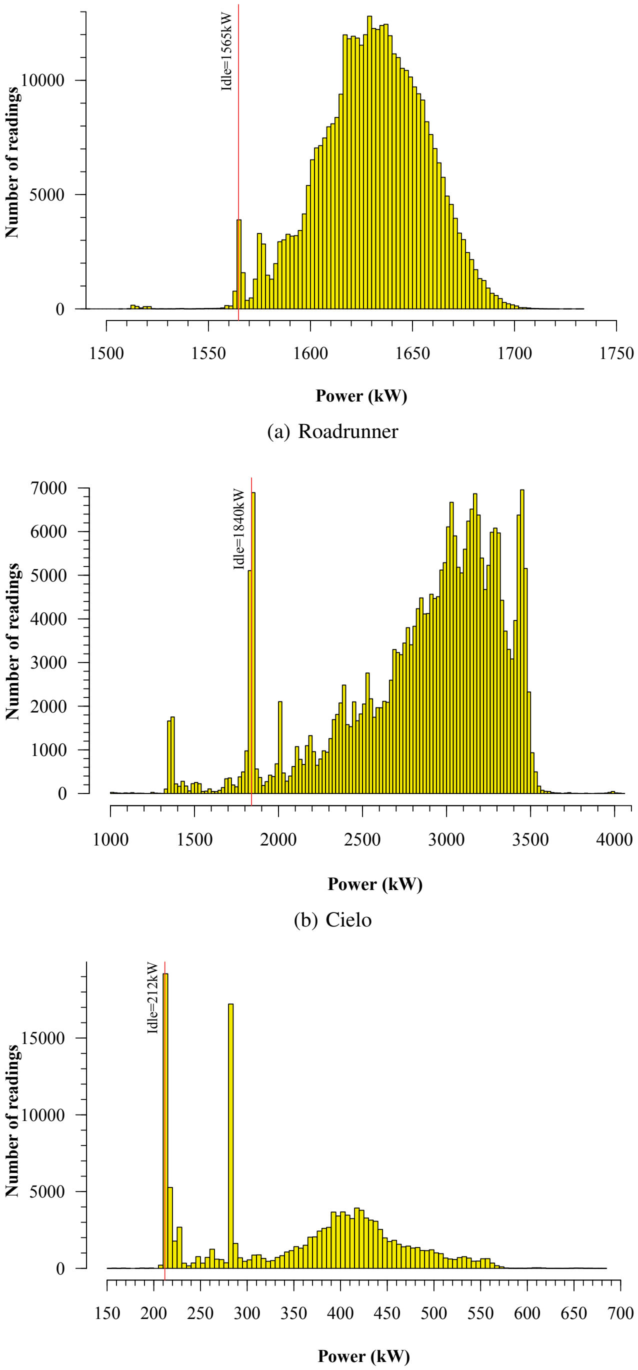
For example, if power were reduced by half, then any concomitant slowdown of less than 2x would result in a saving of energy while a slowdown of more than 2x would result in a squandering of energy.

We begin by determining the idle power for each of our three supercomputers. Figure 4 replots Figure 2 as a histogram, discarding outliers on the left part of the graph that are presumably observed during full-system boot or power down. Note that each figure is plotted with different axes to clarify the shape of the curve; Roadrunner's horizontal range, for example, is extremely narrow. Idle power, drawn with a vertical red line in each histogram, was calculated by selecting the first "significant" rise (defined as 10% of the highest peak) and corroborating that with visual inspection.

Given idle power, we can now treat that as the limit in power saving achievable by throttling processor frequencies and voltages. Table IV presents the data and outcome of our energy calculation. By means of explanation, Total Time represents Figure 2's horizontal range but with gaps elided; Total Energy is the area under the curve in Figure 2; and Idle Power is taken from Figure 4 as described above. Scaled Time, the time that the entire workload would take if run at idle power but the same total energy, is computed as Total Energy ÷ Idle Power. The tolerable slowdown is therefore the quotient of Scaled Time and Total Time (i.e., measured time).

The conclusion one should draw from Table IV is somewhat disappointing: Almost any slowdown in execution speed will result in an increased energy cost for performing the complete workload on any of the three supercomputers. Even Luna, the most tolerant of slowdown of the systems shown in the table, would need to run no more than 48% slower at idle power than at average power in order to observe a net benefit in energy expended.

On a more optimistic note, Roadrunner, Cielo, and Luna, like most supercomputers, always run their CPUs at the highest available clock rate (and therefore power) instead of using a dynamic power governor to raise and lower clock speeds on demand [21], as is common on desktop and laptop systems. Consequently, it may be possible to lower the idle-power levels below those shown in Figure 4 and thereby tolerate a greater performance loss than what is indicated by Table IV.



(a) Roadrunner

(b) Cielo

(c) Luna

Fig. 4: Histograms of supercomputer power usage over the time period from 1JAN2011 to 30APR2012

B. Controlled studies

To complement our measurements of production workloads over a long time frame (Section IV-A) we additionally performed some controlled power studies on Luna. We were granted exclusive access to the entire Luna system for 10 hours and used this time in an attempt to answer the following questions:

- How well do full-system power readings at the switchboard match the aggregate intra-rack power readings?
- Are real applications' power characteristics qualitatively similar or different to those of kernel benchmarks?

Our methodology was as follows. We selected two kernel benchmarks to run—matrix-matrix multiplication (mmult) and the High-Performance LINPACK (HPL) benchmark [18]—and two representative LANL applications—xRAGE, a radiation-hydrodynamics code [22], and SPaSM, a molecular-dynamics

TABLE IV: Maximal tolerable slowdown for a fixed energy budget

System	Total time (s)	Total energy (J)	Idle power (W)	Scaled time (s)	Tolerable slowdown
Roadrunner	3.32×10^7	5.26×10^{13}	1.56×10^6	3.36×10^7	1.01 (1%)
Cielo	3.33×10^7	8.02×10^{13}	1.84×10^6	4.36×10^7	1.31 (31%)
Luna	9.79×10^6	3.08×10^{12}	2.12×10^5	1.45×10^7	1.48 (48%)

code [23]. We began by letting Luna idle to get a consistent idle-power reading. Then we ran each of mmult, HPL, xRAGE, and SPaSM in turn, monitoring power at both the switchboard level and the "shelf" (10-node) level, allowing Luna to return to baseline power between runs. mmult is a single-process program and was run simultaneously on every core of the machine. The rest of the programs were run as 64-node jobs, filling the machine with those. We chose 64 nodes because, as Figure 5 indicates, over 50% of the jobs that run on Roadrunner, Cielo, and Luna utilize 64 or fewer nodes. (Note that this is a measure of job count, not execution time; otherwise the numbers would be quite different.) The cumulative density function (CDF) drawn in the figure excludes single-node jobs, which are assumed to represent compilations, debugging sessions, interactive usage, and other tasks that are not production runs of scientific applications.

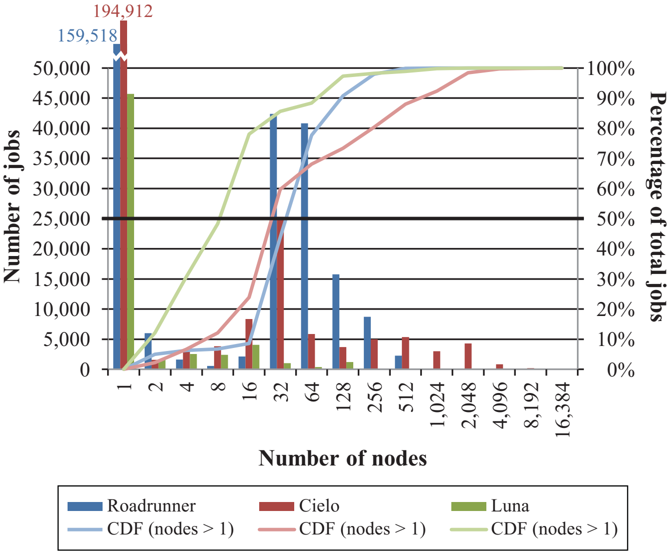


Fig. 5: Histogram of job sizes from 1JAN2011 to 30APR2012

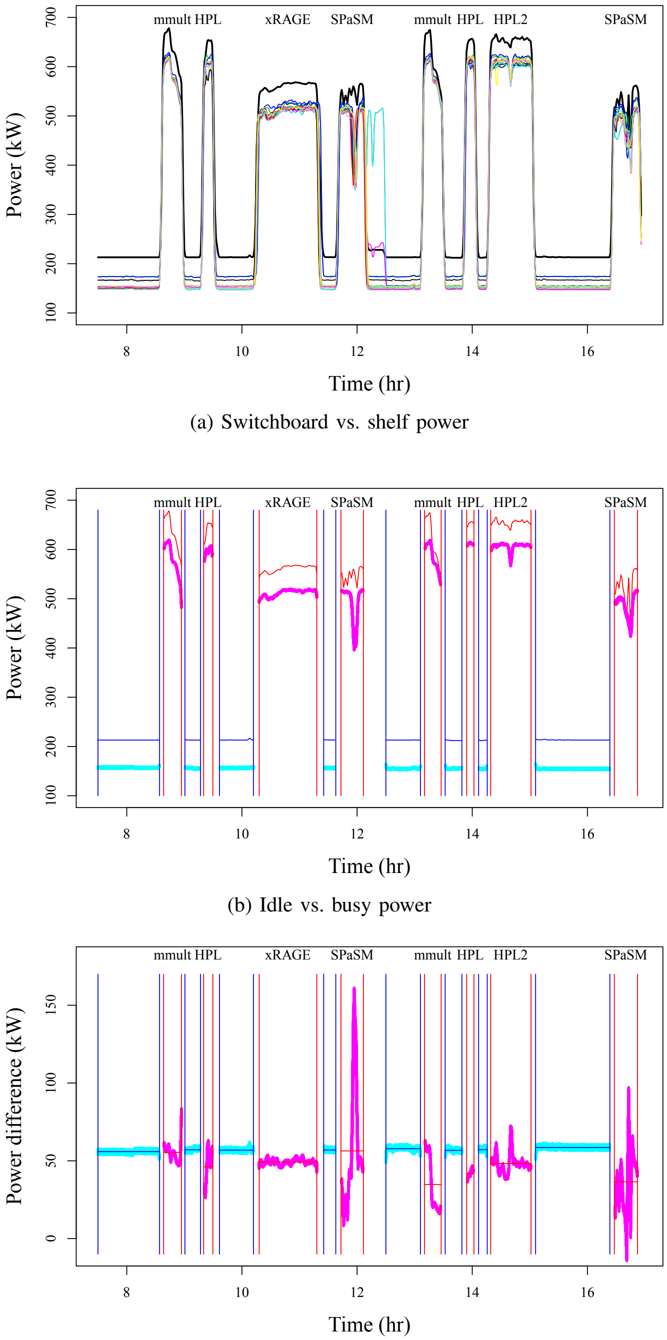
To ensure that our test workload does not leave the system in a different power state from how it began we then re-ran mmult and HPL to confirm that the power readings matched the previous readings. To quantify the impact of our choice of using 64-node jobs we ran a full-system (1,540-node) HPL job. Finally, after a long cooling-down period (partly including a few experiments that failed to launch), we re-ran SPaSM to get a second reading on its power usage.

When analyzing our data after our 10-hour session, we discovered, to our chagrin, that only 15 of the shelf-level power monitors had returned reliable data. Subsequent small-scale experimentation with Luna indicated that polling those monitors too quickly sometimes puts them into an erroneous state. We therefore employed some statistical analysis of the good monitors, described below, to compensate for the missing data. Furthermore, it was not possible to precisely synchronize in time the readings of the shelf power monitors (relative to the speed at which applications can change their power demands) so there is some skew in the readings. Again, we applied rigorous statistical analysis to adapt as well as possible to the available data.

Figure 6 plots the results of our experiments on Luna. In that figure, "HPL" represents the set of 64-node LINPACK jobs, and "HPL2" represents the full-system LINPACK job. Measurements of the 15 good shelf-level power monitors and the aggregate of the switchboard monitors are plotted in Figure 6(a). The thick, black line in the figure represents the switchboard power, and the the thin, colored lines represent per-shelf power. As Figure 6 indicates, the shelf power draws are fairly consistent. The shelf power draw curves were multiplied by 154 so that if every shelf behaved as that particular shelf, it would represent the cumulative power draw from all shelves. Also, these curves are not the raw data; a small amount of kernel smoothing was performed to fill in missing records.

To get a handle on whether the differences between programs seen in Figure 6(a) are statistically real, we performed a one-way analysis of variance (ANOVA) on the maximum power draw during the execution of each code. This leads to largely significant differences between all programs except between HPL2 (the full-system LINPACK) and HPL (LINPACK on each node), and between HPL2 and mmult, both of which are mildly significant differences. SPaSM and xRAGE are also not significantly different in terms of expected maximum power draw.

Because of time alignment issues—the shelf times lag by as much as 30 seconds—it is virtually impossible to recover any kind of useful comparison from switchboard to shelf power on ramp up and cool down (since they happen so rapidly). Therefore, Figure 6(b) plots the comparison between switchboard power draw and a projection of the cumulative shelf power draw during idle times (in between blue bars) and during the middle of code execution (in between red bars). The projection of cumulative shelf power draw was obtained by resampling the 15 shelf curves that were collected to fill in the "missing" 139 shelves to get plausible records for the total of 154 shelves. This can be considered an empirical Bayesian procedure [24] where the distribution of shelf power curves is estimated with the empirical distribution [25] (i.e., each observed curve is given probability 1/15), and the uncertainty in



(a) Switchboard vs. shelf power

(b) Idle vs. busy power

(c) Idle vs. busy power deltas

Fig. 6: Controlled power studies on Luna

the total is then sampled according to the estimated distribution.

This process of sampling the missing shelf power curves, then adding them up produces many plausible curves for the cumulative shelf power draw. 1000 such curves are plotted in Figure 6(b) (in cyan for idle times and magenta for code execution). The band of cyan and magenta curves therefore includes the uncertainty in the cumulative shelf power resulting from our ability to obtain power data for only 15 of the shelves. However, this does *not* include uncertainty related to the time of measurement, which is likely important, especially for rapidly changing (in terms of power usage) codes like SPaSM.

To highlight the differences between the switchboard curve and the plausible cumulative shelf power curves, Figure 6(c) replots the data as deltas between application-execution and idle times. Overall, there is about a 53 kW difference (averaged across all time in the plot and across plausible shelf power curves), with a confidence interval for the average difference over the time of the experiment of $(51.7, 54.4)$ kW (i.e., an accounting of the uncertainty in total shelf power).

It is very difficult to assess the differences between switchboard and shelf power locally in time because of the time lag issues with the shelf-power readings mentioned above. This is particularly problematic for a program like SPaSM, whose power draw fluctuates rapidly during execution. Hence, it is unclear whether the differences between switchboard and shelf power between experimental conditions (including idle condition) in Figure 6(c) are due mostly to program differences or time resolution/accuracy issues. With that in mind, an ANOVA of the max power difference (between switchboard and shelf) during the 16 experimental trials (including the eight code runs and eight idle times as separate trials) found no significant differences between idle, mmult, HPL, xRAGE, and HPL2. However, SPaSM was significantly different from all other programs and from idle. Again, it is unclear how much of this is simply a time-alignment issue.

V. Future Work

Having access to real-world power data on production supercomputers is an invaluable first step towards a variety of potential power studies. For starters, future work ought to correlate power data with job information. LANL maintains vast data on every job ever run on Roadrunner, Cielo, and Luna. Finding ways to link power measurements with job characteristics may result in some interesting insights regarding the way different applications consume power. If these insights lead to a predictive capability, then avenues for future research can include means for full-system power provisioning or power capping, broadening the node-level scope of current approaches such as Intel's Node Manager [20]. This can include job-scheduling aspects such as coscheduling high- and low-power jobs—or even phases within a job (e.g., computation versus file I/O)—to maintain a specified average power.

VI. Conclusions

In this work we presented power measurements taken over a long period of time (16 months) on three large and architecturally disparate supercomputers (two Top10 systems and a Top50 system) running production workloads. To our knowledge, this is the first non-controlled study of power usage of a scientific workload performed at supercomputing scales. The following conclusions can be drawn from the data we presented in this paper:

- Variability in power draw can be quite different across different architectures, even when running a similar mix of applications.

- LANL's scientific workload draws substantially less power on average (70–75%) than the LINPACK benchmark. Consequently, if supercomputing facilities are specced to LINPACK's power needs, a substantial amount of trapped power capacity will be available to the data center.
- Throttling performance to maintain a maximum power level has the potential to succeed without disrupting the majority of applications. An an extreme example, Roadrunner can have its allocated power capped by 40%—even assuming a 20% safety margin below "nameplate" power—without impacting more than 0.3% of the workload that normally runs on that machine at LANL. This translates into an annual savings of over $1.3 million (US$), assuming a power and cooling cost of $1 million per megawatt per year [26].
- A simple statistical analysis of our power data indicates that there may be opportunity for a power-aware job scheduler to reduce the LANL workload's peak power consumption by coscheduling jobs that consume little power alongside jobs that consume substantial power.
- There is so little difference between average and idle power on the three supercomputers studied that the best way to reduce the energy needed to perform a scientific workload is to run at full speed/maximum power rather than trying to reduce power and pay a penalty in execution time.

In summary, we have analyzed the power used by a production supercomputing environment. While our results are unlikely to precisely represent other supercomputing data centers, supercomputing platforms, or workloads, we believe that our methodolgy for analysis can be applied universally to determine salient characteristics about the systems in question and about the potential to achieve greater power efficiencies than what are currently observed.

References

[1] P. Kogge, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K. Hill, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R. Stanley, and K. Yelick, "ExaScale computing study: Technology challenges in achieving exascale systems," Defense Advanced Research Projects Agency (DARPA), Tech. Rep., Sep. 28, 2008. [Online]. Available: http://www.cse.nd.edu/Reports/2008/TR-2008-13.pdf

[2] J. Dongarra, P. Beckman, T. Moore, P. Aerts, G. Aloisio, J.-C. Andre, D. Barkai, J.-Y. Berthou, T. Boku, B. Braunschweig, F. Cappello, B. Chapman, X. Chi, A. Choudhary, S. Dosanjh, T. Dunning, S. Fiore, A. Geist, B. Gropp, R. Harrison, M. Hereld, M. Heroux, A. Hoisie, K. Hotta, Z. Jin, Y. Ishikawa, F. Johnson, S. Kale, R. Kenway, D. Keyes, B. Kramer, J. Labarta, A. Lichnewsky, T. Lippert, B. Lucas, B. Maccabe, S. Matsuoka, P. Messina, P. Michielse, B. Mohr, M. S. Mueller, W. E. Nagel, H. Nakashima, M. E. Papka, D. Reed, M. Sato, E. Seidel, J. Shalf, D. Skinner, M. Snir, T. Sterling, R. Stevens, F. Streitz, B. Sugar, S. Sumimoto, W. Tang, J. Taylor, R. Thakur, A. Trefethen, M. Valero, A. van der Steen, J. Vetter, P. Williams, R. Wisniewski, and K. Yelick, "The International Exascale Software Project roadmap," *International Journal of High Performance Computing Applications*, vol. 25, no. 1, pp. 3–60, Feb. 2011, DOI: 10.1177/1094342010391989, ISSN: 1094-3420.

[3] H. Meuer, E. Strohmaier, H. Simon, and J. Dongarra, "Top500 supercomputer sites: November 2011." [Online]. Available: http://www.top500.org/lists/2011/11

[4] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *Proceedings of the 34th Annual International Symposium on Computer Architecture*, San Diego, California, Jun. 9–13, 2007, pp. 13–23, DOI: 10.1145/1250662.1250665, ISBN: 978-1-59593-706-3.

[5] J. H. Laros III, K. T. Pedretti, S. M. Kelly, W. Shu, and C. T. Vaughan, "Energy based performance tuning for large scale high performance computing systems," in *Proceedings of the 20th High Performance Computing Symposium*, Orlando, Florida, Mar. 26–29, 2012. [Online]. Available: http://www.cs.sandia.gov/~jhlaros/publications/HPC2012_Laros.pdf

[6] A. Rawson, J. Pflueger, and T. Cader, "The Green Grid data center power efficiency metrics: PUE and DCiE," The Greed Grid, White Paper #6, Oct. 23, 2007. [Online]. Available: http://www.thegreengrid.org/~/media/WhitePapers/White_Paper_6_-_PUE_and_DCiE_Eff_Metrics_30_December_2008.pdf

[7] S. Greenberg, E. Mills, B. Tschudi, P. Rumsey, and B. Myatt, "Best practices for data centers: Lessons learned from benchmarking 22 data centers," in *Proceedings of the 2006 ACEEE Summer Study on Energy Efficiency in Buildings*. Pacific Grove, California: American Council for an Energy-Efficient Economy, Aug. 13–18, 2006.

[8] K. Barker, K. Davis, A. Hoisie, D. Kerbyson, M. Lang, S. Pakin, and J. C. Sancho, "Entering the petaflop era: The architecture and performance of Roadrunner," in *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*, Austin, Texas, Nov. 15–21, 2008, DOI: 10.1109/SC.2008.5217926, ISBN: 978-1-4244-2835-9.

[9] J. A. Kahle, M. N. Day, H. P. Hofstee, C. R. Johns, T. R. Maeurer, and D. Shippy, "Introduction to the Cell multiprocessor," *IBM Journal of Research and Development*, vol. 49, no. 4/5, pp. 589–604, Jul./Sep. 2005, DOI: 10.1147/rd.494.0589, ISSN: 0018-8646.

[10] *InfiniBand Architecture Specification Release 1.2.1*, InfiniBand Trade Association, Nov. 2007. [Online]. Available: http://www.infinibandta.org/specs/

[11] W. Feng and H. Lin, "The Green500 list: Year two," in *Proceedings of the Sixth Workshop on High-Performance, Power-Aware Computing, 24th IEEE International Parallel and Distributed Processing Symposium*, Atlanta, Georgia, Apr. 19–23, 2010, DOI: 10.1109/IPDPSW.2010.5470905.

[12] C. Vaughan, M. Rajan, R. Barrett, D. Doerfler, and K. Pedretti, "Investigating the impact of the Cielo Cray XE6 architecture on scientific application codes," in *Proceedings of the Workshop on Large-Scale Parallel Processing, 25th IEEE International Parallel and Distributed Processing Symposium*, Anchorage, Alaska, May 16–20, 2011, pp. 1831–1837, DOI: 10.1109/IPDPS.2011.342, ISSN: 1530-2075.

[13] R. Alverson, D. Roweth, and L. Kaplan, "The Gemini system interconnect," in *Proceedings of the IEEE 18th Annual Symposium on High Performance Interconnects*, Mountain View, California, Aug. 18–19, 2010, pp. 83–87, DOI: 10.1109/HOTI.2010.23.

[14] M. Feldman, "Appro comes up multi-million dollar winner in HPC procurement for NNSA," *HPCwire*, Jun. 8, 2011. [Online]. Available: http://www.hpcwire.com/hpcwire/2011-06-08/appro_comes_up_multi-million_dollar_winner_in_hpc_procurement_for_nnsa.html

[15] National Nuclear Security Administration, "NNSA's Cielo supercomputer approved for classified operation," Washington, DC, Mar. 8, 2011. [Online]. Available: http://www.nnsa.energy.gov/print/mediaroom/pressreleases/cielo3811

[16] G. Aloise, N. Barkakati, and G. C. Wilshusen, "National Nuclear Security Administration needs to improve contingency planning for its classified supercomputing operations," United States Government Accountability Office, Washington, DC, Tech. Rep. GAO-11-67, Dec. 9, 2010. [Online]. Available: http://www.gao.gov/assets/320/313294.pdf

[17] B. Schroeder and G. A. Gibson, "A large-scale study of failures in high-performance computing systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 7, no. 4, pp. 337–351, Oct.–Dec. 2010, DOI: 10.1109/TDSC.2009.4, ISSN: 1545-5971.

[18] J. J. Dongarra, P. Luszczek, and A. Petitet, "The LINPACK benchmark: Past, present and future," *Concurrency and Computation: Practice and Experience*, vol. 15, no. 9, pp. 803–820, Aug. 10, 2003, DOI: 10.1002/cpe.728, ISSN: 1532-0626.

[19] P. Thibodeau, "Los Alamos shuts down supercomputers as fire advances," *Computerworld*, Jun. 29, 2011. [Online]. Available: http://www.computerworld.com/s/article/9218042/Los_Alamos_shuts_down_supercomputers_as_fire_advances_

[20] D. Helman, M. Kelly, and M. Guttmann, "Preserving performance while saving power using Intel Intelligent Power Node Manager

and Intel Data Center Manager," Intel Corp., White Paper, 2009. [Online]. Available: http://www.intel.com/content/dam/doc/white-paper/intelligent-power-node-data-center-manager-paper.pdf

[21] V. Pallipadi and A. Starikovskiy, "The ondemand governor: Past, present, and future," in *Proceedings of the Linux Symposium*, vol. 2, Ottawa, Canada, Jul. 19–22, 2006, pp. 215–230. [Online]. Available: http://www.linuxsymposium.org/2006/linuxsymposium_procv2.pdf

[22] M. Gittings, R. Weaver, M. Clover, T. Betlach, N. Byrne, R. Coker, E. Dendy, R. Hueckstaedt, K. New, W. R. Oakes, D. Ranta1, and R. Stefan, "The RAGE radiation-hydrodynamic code," *Computational Science & Discovery*, vol. 1, no. 1, Oct.–Dec. 2008, DOI: 10.1088/1749-4699/1/1/015005, ISSN: 1749-4699.

[23] S. Swaminarayan, K. Kadau, T. C. Germann, and G. C. Fossum, "369 Tflop/s molecular dynamics simulations on the Roadrunner general-purpose heterogeneous supercomputer," in *Proceedings of the ACM/IEEE SC2008 Conference*. Austin, Texas: IEEE Press, Nov. 15–21, 2008. DOI: 10.1109/SC.2008.5214713.

[24] G. Casella, "An introduction to empirical Bayes data analysis," *The American Statistician*, vol. 39, no. 2, pp. 83–87, May 1985, DOI: 10.2307/2682801, ISSN: 1537-2731.

[25] G. R. Shorack and J. A. Wellner, *Empirical Processes with Applications to Statistics*. New York, New York: Wiley, Apr. 1986. ISBN: 978-0471867258.

[26] W. Feng, X. Feng, and R. Ce, "Green supercomputing comes of age," *IT Professional*, vol. 10, no. 1, pp. 17–23, Jan.–Feb. 2008, DOI: 10.1109/MITP.2008.8, ISSN: 1520-9202.